ILLINOIS

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

PRODUCTION NOTE

University of Illinois at
Urbana-Champaign Library
Large-scale Digitization Project, 2007.

Technical Report No. 545

THE DEVELOPMENT AND TESTING OF MEASURES TO ASSESS SCIENCE CONCEPT AND PROCESS ACQUISITION IN FIRST-, SECOND-, AND THIRD-GRADE STUDENTS

C. Nicholas Hastings
Linda A. Meyer
University of Illinois at Urbana-Champaign
Robert L. Linn
University of Colorado at Boulder

December 1991

Center for the Study of Reading

TECHNICAL REPORTS

College of Education
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
174 Children's Research Center
51 Gerty Drive
Champaign, Illinois 61820

CENTER FOR THE STUDY OF READING

Technical Report No. 545

THE DEVELOPMENT AND TESTING OF MEASURES TO ASSESS SCIENCE CONCEPT AND PROCESS ACQUISITION IN FIRST-, SECOND-, AND THIRD-GRADE STUDENTS

C. Nicholas Hastings
Linda A. Meyer
University of Illinois at Urbana-Champaign
Robert L. Linn
University of Colorado at Boulder

December 1991

University of Illinois at Urbana-Champaign
51 Gerty Drive
Champaign, Illinois 61820

The work upon which this publication was based was supported in part by the Office of Educational Research and Improvement under Cooperative Agreement No. G0087-C1001-90 with the Reading Research and Education Center and in part by Grant No. MDR 85-50320 from the National Science Foundation. The publication does not necessarily reflect the views of the agencies supporting the research.

1991-92 Editorial Advisory Board

Abstract

Whereas there is reason to believe that American elementary school children are taught less science than are children in other countries, there is also reason to believe that the tests used to measure American children's knowledge of science concepts and processes are inadequate. Results from a longitudinal study of science performance that examined science textbooks and curricula were used to develop three criterion-referenced measures, which were administered to approximately 650 first-, second-, and third-grade children. Results of descriptive, correlational, and factor analyses show that while these measures correlate highly with norm-referenced measures of verbal ability and science knowledge, they load strongly together on a separate factor.

THE DEVELOPMENT AND TESTING OF MEASURES TO ASSESS SCIENCE CONCEPT AND PROCESS ACQUISITION IN FIRST-, SECOND-, AND THIRD-GRADE STUDENTS

In the last few years, numerous reports (Doran & Jacobson, 1987; Hueftle, Rokow, & Welch, 1983; Jacobson, Takemura, Doran, Kojima, Humrich, & Miyake, 1986; National Assessment of Educational Progress, 1988; Raizen & Jones, 1985) have pointed to the overall lower achievement of American students in science, particularly in the physical sciences, when compared to American students of the last two decades and to students from other countries.

As part of the effort to determine why American students do so poorly on science measures, researchers began to investigate the adequacy of the measures themselves (Jones, 1989). These investigations have revealed problems with the measures and have led to calls for the development of better indicators of student performance. Furthermore, as other investigations have begun to show that science knowledge is "clearly more school-dependent than reading comprehension" (Zuzovsky & Tamir, 1989), others have called for the development of measures that are more sensitive to instruction.

As part of a longitudinal study of science concept acquisition, we examined existing measures of science concept acquisition. We found that the instruments for the lower elementary grades required children to look at pictures or read substantial amounts of text. We also found that a broad range of concepts was represented in each instrument, thereby suggesting that the measures do not focus on content unique to each grade level. As a result of our examination of the existing measures, we decided to develop our own battery of measures of science concept acquisition. In this report, we describe the battery of tests we developed for use at the first-, second-, and third-grade levels. We then present the results of our administration of these tests, along with findings from our administration of norm-referenced tests to comparable groups of students in three school districts participating in the longitudinal study.

Characteristics of Existing Measures

Three nationally normed science tests are commonly used in the lower elementary grades: the Tests of Basic Experiences (Moss, 1978); the Sequential Test of Educational Progress--science subtest (Educational Testing Service, 1979); and the Comprehensive Tests of Basic Skills (CTB McGraw/Hill, 1981). Each of these tests requires children to look at pictures or read sentences or passages. Each test also covers many different scientific concepts.

Tests of Basic Experiences. Our first step in developing new instruments was to study science tests already available for the lower elementary grades. We found just one test to measure science knowledge at the kindergarten level--the Tests of Basic Experiences-2 (TOBE-2). According to its developer, this test deals with "environmental awareness and . . . [is] based on the observation and processing of events in the students' daily lives. For example, a student can observe which objects do and do not float and then make appropriate generalizations based on classification of objects" (Moss, 1978, p. 2). The 26 items on the science test are classified into eight topics such as "plant life," "animal behavior and characteristics," and "force, motion, and mechanics." Two typical items from the TOBE-2 appear in Figure 1.

[Insert Figure 1 about here.]

The first item in the figure shows four line drawings that include shadows. The second item has four illustrations of children moving a large box. In both cases, the illustrations are clear, and they suggest single concepts. An examiner reads the items to the children and tells them to mark the picture that shows what the item is about. In the first item, they are told to mark the picture where the shadow is

right for the object. In the second item, they are told to mark the picture of the box that could be moved most easily.

Sequential Test of Educational Progress--science subtest. As Figure 2 illustrates, the items first-, second-, and third-grade children encounter on norm-referenced science tests such as the science subtest of the Sequential Test of Educational Progress (STEP) are quite different from those on the TOBE-2.

[Insert Figure 2 about here.]

Most of the 50 items on the STEP are similar to these shown. Each item has written instructions that students must read to themselves. Children cannot easily look at the illustrations and have a sense of what the item is about. Some items (see Figure 3) on this test are completely dependent upon the students' reading ability.

[Insert Figure 3 about here.]

Comprehensive Tests of Basic Skills. Another widely used test is the Comprehensive Tests of Basic Skills (CTBS). The CTBS has much to recommend it. It covers a variety of content areas, and it is easily administered. Its readability appears to be at the appropriate grade levels. Yet, even on this test, the Level E science test score has a much higher correlation with a student's total reading test score (.77) than was obtained when students were not required to read the test items on Level D of the CTBS (.52) (CTB/McGraw-Hill, 1982). A correlation as high as .77 suggests that there is relatively little reliable variance on the science test that is not common with reading comprehension ability.

Other tests. Science tests are included in most of the major achievement test batteries for use with students in the upper elementary grades through high school. Some of the batteries also include a test of science in the levels of tests designed for children in the primary grades, usually starting in second grade. The science tests for early grades typically depend heavily on students' general knowledge about their environment. Frequently, items on such tests can be answered by a child who knows the meaning of a word. Hence, on some elementary science tests, vocabulary knowledge may be more important than knowledge of the underlying concepts or the process of scientific reasoning.

Frank (1978) classified a total of 765 items from 12 standardized science tests using the first four developmental levels of Bloom's (1956) taxonomy. He placed only 2% (36 out of 765) of the items in the two higher categories: application and analysis. The majority (78%) of the items were placed in the knowledge category, the lowest of Bloom's first four levels of intellectual development, and the remaining 20% were placed in the comprehension category. Morgenstern and Renner (1984) found a similar emphasis on factual knowledge in their analysis of 12 science tests. Fully 90% of the items were placed in the "recall" category.

The development of factual knowledge is an appropriate objective of the elementary science curriculum. However, items eliciting factual knowledge cannot be expected to assess the development of scientific concepts, the ability to apply concepts, or the process of making generalizations or drawing inferences from observations. Even the few items on standardized tests that are classified as inferential reasoning or evaluation fall short of adequately assessing these higher level objectives. This is partially due to the limitations imposed by group-administered, multiple-choice test items, especially those that do not require reading.

The scarcity of questions that require students to apply concepts or that get at the processes of analysis or evaluation is a serious limitation of standardized science tests. So, too, is the confounding of reading ability with the assessment of a student's understanding of science concepts and principles. Yet a third limitation is the result of the necessarily generalized character of a national, standardized test. Such tests are designed to be applicable to a wide range of curriculum materials and instructional sequences. As a consequence, the overlap of a test with the content of specific curriculum materials used in a

particular classroom and with content covered by a particular teacher can be quite variable (Leinhardt, 1983; Schmidt, 1983).

The criterion-referenced tests we have developed differ from available standardized science tests in three important respects: (a) they emphasize application, analysis, and evaluation; (b) they minimize the dependence upon reading ability; and (c) they contain items made up from content that overlapped in the school districts participating in this study. In other words, the items reflect content common to the three school districts.

Method

The Setting

Three school districts in the midwest have participated in our longitudinal study. In two of these districts, every teacher and child at the appropriate grade levels has taken part in the study, while in the third district, all children at the appropriate grade levels in one school have been involved. We followed two cohorts of children in each district. Cohort 1 consisted of children who entered kindergarten in 1983, Cohort 2 of children who entered in 1984. The districts represented a variety of geographic and cultural settings and utilized several different instructional approaches.

District A is in a somewhat self-contained small town in the center of the state. In this district, there are four kindergarten classes, four first grades, and three second grades in one elementary school. This district is well known for its high student performance in reading comprehension in first, second, and third grades, and for its average student performance in science.

District B is in a small town that is about a 25-minute drive from the larger community in which many of its citizens work. The district had seven classes at each grade level for Cohort 1 and six classes per level for Cohort 2. All of the students in this district attend the same elementary school. The district is known for average student performance in reading in the lower grades and high student performance in reading in the middle grades. It is also known for above-average performance in science throughout its system.

District C bears some resemblance to an urban setting because of the ethnic diversity of its student population. One elementary school participated from this district. The children are of mixed backgrounds. Black, Hispanic, and White children attend the school from this district participating in the study. There are nine other elementary schools in this district. Bilingual children receive instruction in Spanish as well as English. They are known for average performance in reading and science.

Test Development

In general, we used the following procedures to develop tests to assess students' concept acquisition in three different science content domains. First, we analyzed curriculum materials and curriculum-embedded tests used in the three school districts, and information obtained from classroom observations and teacher interviews to determine content areas that were either common to all the districts, specific to one district, or not taught within any of the districts. From this analysis, we selected three subject domains: (a) plants, which was common to all districts and which was taught from the first grade through the third grade; (b) three forms of matter, which was common to all three districts and which was introduced during the second grade and continued for at least two years; and (c) motion, which was not introduced in any of the districts' curricula until the fourth grade and which, therefore, had not been formally taught to any of the children in our study by third grade.

We further analyzed these content domains to determine the concepts, processes, and vocabulary introduced within each curriculum by grade level. Items were then developed, piloted, and revised to

assess both on-level and out-of-level concepts or processes within each of the three content domains. On-level items were those that required information that had already been covered by the children, while out-of-level or "extension" items dealt with material that was to be introduced to the students at a later time in their studies. These procedures yielded four tests: the Error Detection test, the Plants test, the Three Forms of Matter test, and the Motion test.

The Error Detection Test

The Error Detection test (ED) is a multipurpose instrument for use with first graders to obtain information about their reading abilities as well as their acquisition of concepts within the content domain of plants. The measure is intended to be individually administered and contains two subsets of items. The first is the Absurd Target Word subtest, which consists of a series of 10 sentences or sets of sentences that contain a conceptually incorrect word, for example, "The cookies will bloom in the spring." Here, either the word *cookies* or the word *bloom* is inappropriate. Children are directed to read the sentences aloud and to identify the word that "spoils the meaning." Decoding errors are recorded for reading diagnostic purposes, but to minimize the effects of low decoding abilities, the errors are verbally corrected for the child by the examiner. Children receive 1 point for correctly identifying the inappropriate word, and they are then asked to explain or support their rationale in making that choice. The support is then scored as appropriate or inappropriate. Thus, children receive three scores for their performance on the Absurd Target Word section of the test: the number of decoding errors (DAW), the number of correct identifications (IAW), and the number of correct statements supporting their choices (SAW).

The second portion of the ED is the Impossible Sequence subtest, which consists of six "impossible sequences." A typical item reads "Put a carrot top in a pan of water. First you will see new leaves. Then you will see a root." Children are asked to read aloud each sequence, to identify "what happened at the wrong time," and to present support for their choice. Decoding errors (DIS) are recorded and corrected, and points are given for appropriate identifications (IIS) and for giving a rationale in support of their identification of an impossible sequence (SIS).

Plants Test

The Plants test (PTS) contains both on-level and out-of-level items based on the content domain of plants. It is intended for group administration to second graders. The students are presented 20 items, 4 of which have more than one correct response, for a total of 33 possible points.

To reduce reading effects, items on the PTS are presented in the form of line illustrations, and the examiner gives all directions orally (see Figure 4). The items vary along two dimensions: their degree of level appropriateness and their degree of emphasis on factual knowledge versus process-concept application. Along the first dimension, the majority of items are based on information on which the children would have received formal instruction by the end of second grade. For example, plant parts (roots, leaves, stems, and flowers) and to some extent their functions are presented during second grade in all three districts.

[Insert Figure 4 about here.]

However, eight items are included that require knowledge about topics that are not presented until third grade in any of the schools participating in this study (e.g., photosynthesis). The second dimension on which item content varies is the degree of emphasis on factual knowledge versus process-concept application. We attempted to include a large number of process-concept oriented items during our test development procedure. An example of a process-content item might be the "part that collects water."

Three Forms of Matter

The Three Forms of Matter (3FM) test consists of 34 dichotomously scored items and is group administered. Again, to control for reading effects, all directions are read by the examiner to the children. When item stems and/or alternatives are presented in written form (only 10 of the 34 items are line drawings), the stems and alternatives are also read aloud. For example, one set of items consists of a list of words followed by the letters S, L, and G representing solid, liquid, and gas. Children are read the word (e.g., *oxygen*) and then instructed to "Circle S if oxygen is a solid, L if oxygen is a liquid, or G if oxygen is a gas."

As in the PTS, we selected the majority of items on the 3FM from on-level content that was presented in varying degrees in each of the three districts by the end of second grade. Also, we attempted to emphasize process-concept knowledge as opposed to factual knowledge. Thirteen of the items can be classified as factual information items such as the item asking if oxygen is a solid, liquid, or gas. Twenty-one of the items deal with more conceptual information. For example, one item attempts to assess children's understanding of the concept that evaporation rate is partially determined by surface area (see Figure 5). Children are presented with four line drawings of jars containing water. Two of the jars are lidded, prohibiting evaporation, and the other two are open. One of the open jars has a much larger water surface area and mouth than the other. The students are asked to circle the number of the picture of the container of water that will evaporate fastest.

[Insert Figure 5 about here.]

Motion Test

The last of the four science tests that we developed and administered is the Motion test (MT). All of the items on the instrument may be considered to be out of level because motion is not taught in these districts until at least fourth grade. The MT was included in an end-of-second-grade science test battery to provide baseline information about students' understanding of a content domain that they will eventually study. We hope that future administrations of this instrument will be useful in documenting students' growth within this content domain.

Like the PTS and the 3FM, the MT test can be group administered. Again, the examiner reads the directions aloud. Only 1 of the 20 MT items in the student test booklets presents alternatives in written form, and these alternatives are also read by the examiner. The rest of the items are presented as line drawings. For example, three items depict an overhead view of a person rolling a ball towards a wall from different angles. Students are asked to choose one of three possible paths that the ball will take after it hits the wall (see Figure 6).

[Insert Figure 6 about here.]

Eight of the 20 items might be classified as simple identification items. They consist of a series of drawings of two lines intersecting at either right angles, "straight Ts," or at slanted angles (see Figure 7). The examiner draws examples of "straight Ts" and "non-straight Ts" on the board and asks children to "draw a circle around every T that is a 'straight T.'"

[Insert Figure 7 about here.]

Standardized Tests

We also administered the following standardized tests to allow us to compare the results of the tests we developed.

CIRCUS Reading Test. The CIRCUS Reading test, Level D (Educational Testing Service, 1976) was given to Cohort 1 children in the spring of their second-grade year. This is a relatively traditional group-administered reading test. It is composed of short passages and comprehension questions.

Degrees of Reading Power Test. The Degrees of Reading Power test (DRP) (College Board, 1979) Form PA-8 was administered out of level at the end of the second-grade year to all students. This test involves several passages. Each passage is 5-7 paragraphs long. Each selection has seven cloze blanks. The cloze blanks are purported to be understood only in the context of the preceding and following sentences. The passages increase in difficulty, and children have as long as they need to complete the test, although they may reach passages that are so difficult for them that they stop working and consider themselves finished.

STEP Science. The science subtest of the STEP was administered to Cohort 1 children in the spring of third grade. Students must read items silently to complete this test. It is composed of line drawings and comprehension questions.

TOBE-2. Level K and Level L of the TOBE-2 were used as end-of-year dependent variables for kindergarten and first grade. On these instruments, children are asked to choose one of four line drawings in response to item stems administered orally.

Wide Range Achievement Test. The reading subtest, Level I of the Wide Range Achievement Test (WRAT) (Jastak, Jastak, & Bijou, 1978) was administered at least once a year. Items on the WRAT consist of a series of increasingly difficult words that children read aloud to an examiner. The measure is individually administered and has a stopping rule whereby 12 consecutive errors terminate administration.

Descriptive Results

Error Detection Test. Cohort 1 took the ED twice, once at the end of first grade, and again at the beginning of second grade. Descriptive statistical data for both testings are presented in Table 1 and Table 2 for the total sample and for each of the districts.

[Insert Tables 1 and 2 about here.]

As the tables show, reliability results, computed as coefficient-α, tend to be highest for the decoding subscores, DAW and DIS, for total sample and each district. Also, these α's are relatively consistent from the spring (first grade) testing to the fall (second grade) testing. The correlations between these two measures were high (above .90) for both spring and fall administrations and .73 to .80 for spring-fall comparisons. Rank orderings of the districts on decoding subscores agree with differences found on other reading measures used in the study.

Reliabilities for the more content domain-related subscores were considerably lower. This is especially so for the second-grade administration. A glance at the restriction of range for these scores (e.g., in District B, all children correctly identified 4 of the 6 impossible sequence items) explains the lack of variance. Items were not difficult enough for the children, especially at the second-grade level. For example, only 4 of the 319 students were unable to correctly identify the inappropriate sequence "Watch a plant grow. First you see a little plant. Then you see a bigger plant. Then you see only dirt."

For the first-grade administration, correlations between decoding subscores and process subscores ranged from -.34 to -.50 (e.g., $r_{DAW\,SAW}$ = -.50), while correlations between process scores on the two different subtests ranged from .48 to .60 (e.g., $r_{LAW\,IIS}$ = .48). The relatively high correlations of decoding subscores with process subscores suggests that reading and/or verbal abilities affect performance on the more process-oriented subscores. This is not particularly surprising. Rank orderings of the districts on

process subscores follows the same pattern as rank orderings obtained with several standardized science tests we have administered that show relatively high reading dependence.

Plants Test. This test was administered in the spring. Table 3 presents descriptive statistics for the total sample and for each of the districts for the PTS. Children in District B performed best on this measure; District A had the worst showing. Response patterns to out-of-level items were predictably lower than to on-level items. For example, one item presents the student with a picture of a plant (see Figure 4), and the children are asked to identify the part of the plant "that makes its food." This item had a *p*-value of only .23, as contrasted with the *p*-value of .62 for an item that directs the students to identify the part of the same plant "that collects water."

[Insert Table 3 about here.]

This item may be contrasted with an item having a similar illustration and asking the student to identify "the root." The *p*-value of the "collects water" item was .62, while the "root" identification item was considerably easier having a *p*-value of .96.

In general, midrange *p*-values were obtained for process-concept-oriented items that addressed "common core" on-level content information. It is likely that the PTS's sensitivity to between-school and between-classroom differences may be attributed to the strong emphasis that was placed on such items during test development.

Three Forms of Matter Test. This test was given to Cohort 1 at the end of the second-grade year. Descriptive statistics for total sample and by district are presented in Table 4. Like the Plants test, 3FM showed significant differences between districts and classrooms after controlling for differences in achievement on the TOBE-2 given at the end of kindergarten. Districts were ranked in the same order as they were on the PTS; District B followed by C followed by A.

[Insert Table 4 about here.]

Like the PTS test, the 3FM test reveals between-school differences, whereas the STEP, a standardized test, does not. We feel that this fact is due to our ability to control item content with respect to information actually presented to children in the study and to the PTS's emphasis on science concepts and processes as compared to the STEP's emphasis on facts and vocabulary.

The Motion Test. Table 5 presents descriptive statistical information regarding the Motion test for the total sample and each of the three districts. Although district means are in the same order as in the PTS and 3FM, these differences were found to be non-significant after controlling for science achievement measured by the TOBE-2. Note that coefficient-α's are lower on this measure (.38 to .54 across districts) than on either the PTS or 3FM.

These results are not surprising given that the content domain of motion is not formally introduced in any of the three districts until the fourth grade.

[Insert Table 5 about here.]

P-values for these items were quite high (.87 to .98) as compared to the *p*-values of concept-oriented items on the test that were near or even below chance levels.

These results may confirm that most of these children had not acquired conceptual knowledge about the domain of motion by the end of second grade. We hope that future administrations of the MT will be useful in documenting students' conceptual development about the content domain of motion when that topic is introduced.

Correlational Analyses

Table 6 presents results from a correlational analysis of PTS, 3FM, and MT with three standardized reading tests, the WRAT, Level 1; the CIRCUS, Level D; and the DRP, Form PA-8, and two standardized science measures, the TOBE-2 and the STEP. The WRAT and TOBE-2 were both given during the students' kindergarten year and are thus premeasures. The other six measures were administered in the spring of the second-grade year and are postmeasures.

Examination of Table 6 reveals several potentially interesting results. First, MT does not seem to correlate strongly with any of the other measures. This might have been expected; the majority of items in the instrument present information that has not yet been introduced to the children. It therefore attempts to measure abilities that are undeveloped in the student sample by the end of second grade. A measure of "nonexistent" concepts or abilities should not be expected to correlate highly with measures that tap "real" student abilities.

Our other two second-grade science measures, PTS and 3FM, correlate moderately well with the STEP and each other (r_{PS} = .51, r_{3S} = .60, and r_{P3} = .49). Some degree of overlap between a test of "general" science knowledge such as the STEP, and our more specifically oriented measures is desirable. So, too, is the fact that our measures correlate more highly with a "general science" test than they do with concurrently administered "reading" instruments.

[Insert Table 6 about here.]

The picture obtained from Table 6 is not as clear as we might wish, however. A high degree of "reading-verbal" ability appears to be involved with student performance on the standardized STEP test. The highest correlations in the table are for the STEP with the CIRCUS followed by the STEP with the DRP, r_{SC} = .79 and r_{SD} = .70. The correlation of the two second-grade reading tests with each other is even slightly lower than these (r_{CD} = .67). In fact, the STEP correlates more highly with the WRAT (r_{SW} = .52), a kindergarten reading measure, than it does with the TOBE-2 (r_{ST} = .44), a kindergarten science measure in which all directions are read to the children and all items are pictorially presented. Thus, caution might be in order when comparing our measures with the STEP.

First-order partial correlations controlling for the "reading-verbal" abilities measured by the various reading tests help to clarify the picture somewhat. For example, when the reading comprehension ability measured by the CIRCUS is partialled out, $r_{P3 \cdot C}$ = .38, $r_{PS \cdot C}$ = .39, $r_{3S \cdot C}$ = .34, $r_{ST \cdot C}$ = .38, $r_{PT \cdot C}$ = .30, and $r_{3T \cdot C}$ = .19. Very similar results are obtained when the "reading-verbal" ability in common with the DRP is controlled for. Partial correlation patterns for the STEP show that it continues to have as much or more in common with the reading tests as it does with other science tests (e.g., $r_{SC \cdot D}$ = .60, $r_{SD \cdot C}$ = .38). The 3FM correlates moderately well with the PTS and the STEP but not particularly well with the TOBE-2. Also, partial correlations of the 3FM with reading tests are slightly higher than its correlation with the TOBE-2 when the effects of "verbal-reading" ability are partialled out. The patterns of partial correlations of the PTS with other measures are more in keeping with what might be desirable in a specialized science test. It correlates moderately well with other science measures but does not seem to have much in common with the reading tests (e.g., $r_{PC \cdot D}$ = .14).

Factor Analysis

Overall, correlational analyses can be expected to show only part of the picture at this point in time. A factor analysis was run on these correlational data using a promax rotation. As can be seen in Table 7, although the two factors correlate fairly highly with each other (r_{F1F2} = 0.65), tests that require reading/verbal abilities tend to load rather strongly on the first factor, while our three tests and the TOBE-2 load more heavily on the second factor. We hope that future administrations of "extension

measures" within these specific content domains will provide information for more detailed construct and predictive validation of our measures.

[Insert Table 7 about here.]

Discussion

The development, administration, and analyses of instruments tied to specific content domains within the field of science create a unique opportunity for assessing science concept acquisition. While the instruments we developed are not totally free of the variance in scores associated with "general ability" that is found in standardized science tests, they do provide a means of assessing students' development in content domains that have been covered in varying degrees in their schools, and they are free of the problems associated with instruments that require children to read. Thus these instruments may yield stronger effects between school curricula, teacher instruction, and student scores. Therefore, moderate relationships between performance on these measures and performance on standardized tests of general science knowledge may represent valid proxies for what these second grade students conceptually understand about processes within these specific science content domains.

It is particularly promising that the three custom measures of science learning that did not require the children to read load together to form a factor that we can tentatively identify as science ability. In contrast, the reading measures and the more traditional science tests form a second factor that can aptly be identified as representing general ability. These results in particular suggest that these measures are, in fact, doing what we had hoped. They are measuring student performance in ways that the traditional tests do not while at the same time being moderately correlated to them.

We suggest that further investigations and refinement of these instruments is in order. We also believe that new instruments should be developed at higher grade levels that will continue the traditions of development found in these measures. Such a battery of elementary grade science tests would then offer school districts opportunities for assessing students' ability in science in ways that focus upon science concepts and processes without penalizing them for their reading ability.

References

Bloom, B. S. (Ed.) (1956). *Taxonomy of educational objectives.* Ann Arbor, MI: Edwards Bros.

College Board. (1979). *Degrees of reading power.* New York.

CTB/McGraw-Hill. (1978). *Tests of basic experiences - (TOBE-2).* Monterey Park, CA.

CTB/McGraw-Hill. (1981). *The comprehensive tests of basic skills, form V.* Monterey Park, CA.

CTB/McGraw-Hill. (1982b). *CTBS form V and V preliminary technical manual.* Monterey Park, CA.

Doran, R. L., & Jacobson, W. J. (1987, April). *Preliminary results for the USA participation in the IEA study--1986.* Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Educational Testing Service. (1976). *CIRCUS reading test.* Menlo Park, CA: Addison-Wesley.

Educational Testing Service. (1979). *Sequential test of educational progress Science subtest.* Menlo Park, CA: Addison-Wesley.

Frank, H. J. (1978). An examination of the levels of questions on standardized tests of elementary science. *Science and Children, 14,* 30-32.

Hueftle, S. J., Rakow, S. J., & Welch, W. W. (1983). *Images of science: A summary of results from the 1981-1982 national assessment in science.* Minneapolis: University of Minnesota, Minnesota Research and Evaluation Center.

Jacobson, W. J., Takemura, S., Doran, R. L., Kojima, S., Humrich, E., & Miyake, M. (1986). *Analysis and comparisons of science curricula in Japan and the United States.* New York: Teachers College Press.

Jastak, J. F., Jastak, S. R., & Bijou, S. W. (1978). *Wide range achievement test, level 1.* Wilmington, DE: Jastak Associates.

Jones, L. V. (1989). School achievement trends in mathematics and science, and what can be done to improve them. In E. Z. Rothkopf (Ed.), *Review of research in education* (pp. 307-341). Washington, DC: American Educational Research Association.

Leinhardt, G. (1983). Overlap: Testing whether it is taught. In G. F. Madous (Ed.), *The courts, validity, and minimum competency testing* (pp. 153-170). Boston: Kluwer-Nighoff.

Morgenstern, C. F., & Renner, J. W. (1984). Measuring thinking with standardized tests. *Journal of Research in Science Teaching, 21,* 639-648.

Moss, M. H. (1978). *Norms and technical data book: Tests of basic experiences.* Monterey Park, CA: CTB/McGraw-Hill.

National Assessment of Educational Progress. (1988). *Three assessments of science, 1986: Technical summary.* Denver, CO: Education Commission of the States.

Raizen, S. A., & Jones, L. V. (Eds.). (1985). *Indicators of precollege education in science and mathematics: A preliminary review.* Washington, DC: National Academy Press.

Schmidt, W. H. (1983). Content biases in achievement tests. *Journal of Educational Measurement, 20,* 165-178.

Zuzovsky, R., & Tamir, P. (1989). Home and school contributions to science achievement in elementary schools in Israel. *Journal of Research in Science Teaching, 26,* 703-714.

Table 1

Error Detection Test
Spring 1985 Administration
Descriptive Statistics

Sample	Subtest	α	*N*	$\bar{X}$	Standard Deviation	Range Minimum (0)	Range Maximum (107/116)
Total	DAW	0.97	320	11.90	15.14	0	79
	IAW	0.79	320	8.41	1.88	0	10
	SAW	0.79	320	8.19	2.03	0	10
	DIS	0.95	318	12.56	15.58	0	86
	IIS	0.59	318	5.07	1.14	0	6
	SIS	0.63	316	4.74	1.36	0	6
District A	DAW	0.97	88	7.56	11.49	0	77
	IAW	0.63	88	8.41	1.39	3	10
	SAW	0.80	88	8.07	2.04	0	10
	DIS	0.94	88	8.08	10.42	0	83
	IIS	0.36	88	5.11	0.96	2	6
	SIS	0.68	88	4.77	1.44	0	6
District B	DAW	0.96	153	11.61	11.32	0	76
	IAW	0.74	153	8.68	1.61	2	10
	SAW	0.69	153	8.50	1.65	3	10
	DIS	0.92	153	12.29	11.38	0	83
	IIS	0.59	153	5.13	1.10	0	6
	SIS	0.56	152	4.82	1.24	0	6
District C	DAW	0.98	78	17.46	22.23	0	79
	IAW	0.88	78	7.85	2.60	0	10
	SAW	0.85	78	7.65	2.54	0	10
	DIS	0.92	76	18.45	24.02	0	86
	IIS	0.71	76	4.87	1.40	0	6
	SIS	0.67	75	4.52	1.46	0	6

Table 2

Error Detection Test
Fall 1985 Administration
Descriptive Statistics

Sample	Subtest	α	*N*	$\bar{X}$	Standard Deviation	Range Minimum (0)	Range Maximum (107/116)
Total	DAW	0.98	319	10.91	15.24	0	77
	IAW	0.74	319	8.34	1.67	0	10
	SAW	0.74	319	8.10	1.84	0	10
	DIS	0.95	319	12.30	15.95	0	83
	IIS	0.46	319	5.33	0.88	0	6
	SIS	0.48	319	5.06	1.05	0	6
District A	DAW	0.97	84	7.71	13.07	0	77
	IAW	0.69	84	8.23	1.54	2	10
	SAW	0.65	84	8.06	1.61	0	10
	DIS	0.95	84	9.52	13.29	0	83
	IIS	0.03	84	5.42	0.68	3	6
	SIS	0.23	84	5.18	0.88	3	6
District B	DAW	0.96	154	10.17	12.06	0	77
	IAW	0.61	154	8.77	1.29	3	10
	SAW	0.66	154	8.52	1.51	3	10
	DIS	0.95	154	11.56	13.31	9	83
	IIS	-0.02	154	5.46	0.61	4	6
	SIS	0.08	154	5.20	0.77	2	6
District C	DAW	0.98	81	15.68	20.82	0	77
	IAW	0.82	81	7.65	2.15	0	10
	SAW	0.82	81	7.33	2.35	0	10
	DIS	0.96	81	16.59	21.48	0	83
	IIS	0.75	81	5.01	1.32	0	6
	SIS	0.76	81	4.67	1.48	0	6

Table 3

Plants Test
Descriptive Statistics

Sample	N	α	$\bar{X}$	Standard Deviation	Range	
					Minimum (0)	Maximum (33)
Total	305	0.66	21.70	3.63	12	29
District A	78	0.64	20.73	3.53	12	29
District B	144	0.53	22.50	2.88	14	29
District C	83	0.76	21.22	4.51	12	29

Table 4

Three Forms of Matter Test Descriptive Statistics

Sample	N	α	$\bar{X}$	Standard Deviation	Range	
					Minimum (0)	Maximum (34)
Total	306	0.77	20.34	5.24	5	32
District A	77	0.76	18.95	5.09	5	27
District B	144	0.71	21.27	4.61	10	30
District C	85	0.83	20.02	6.06	6	32

Table 5

Motion Test
Descriptive Statistics

Sample	N	α	$\bar{X}$	Standard Deviation	Range	
					Minimum (0)	Maximum (20)
Total	306	0.47	11.67	2.34	5	18
District A	306	0.47	11.67	2.34	5	18
District B	144	0.38	11.87	2.10	6	18
District C	85	0.52	11.62	2.41	6	18

Table 6

Correlations of PTS, 3FM, and MT with Pre- and Poststandardized Tests

	Standardized Reading Tests			Standardized Science Tests		Customized Science Tests		
	WRAT	CIRCUS	DRP	TOBE-2	STEP	PTS	3FM	MT
WRAT	1.00							
CIRCUS	.45	1.00						
DRP	.44	.67	1.00					
TOBE-2	.32	.39	.31	1.00				
STEP	.52	.79	.70	.44	1.00			
PTS	.34	.36	.39	.40	.51	1.00		
3FM	.44	.54	.54	.36	.60	.49	1.00	
MT	.10	.17	.21	.31	.26	.19	.25	1.00

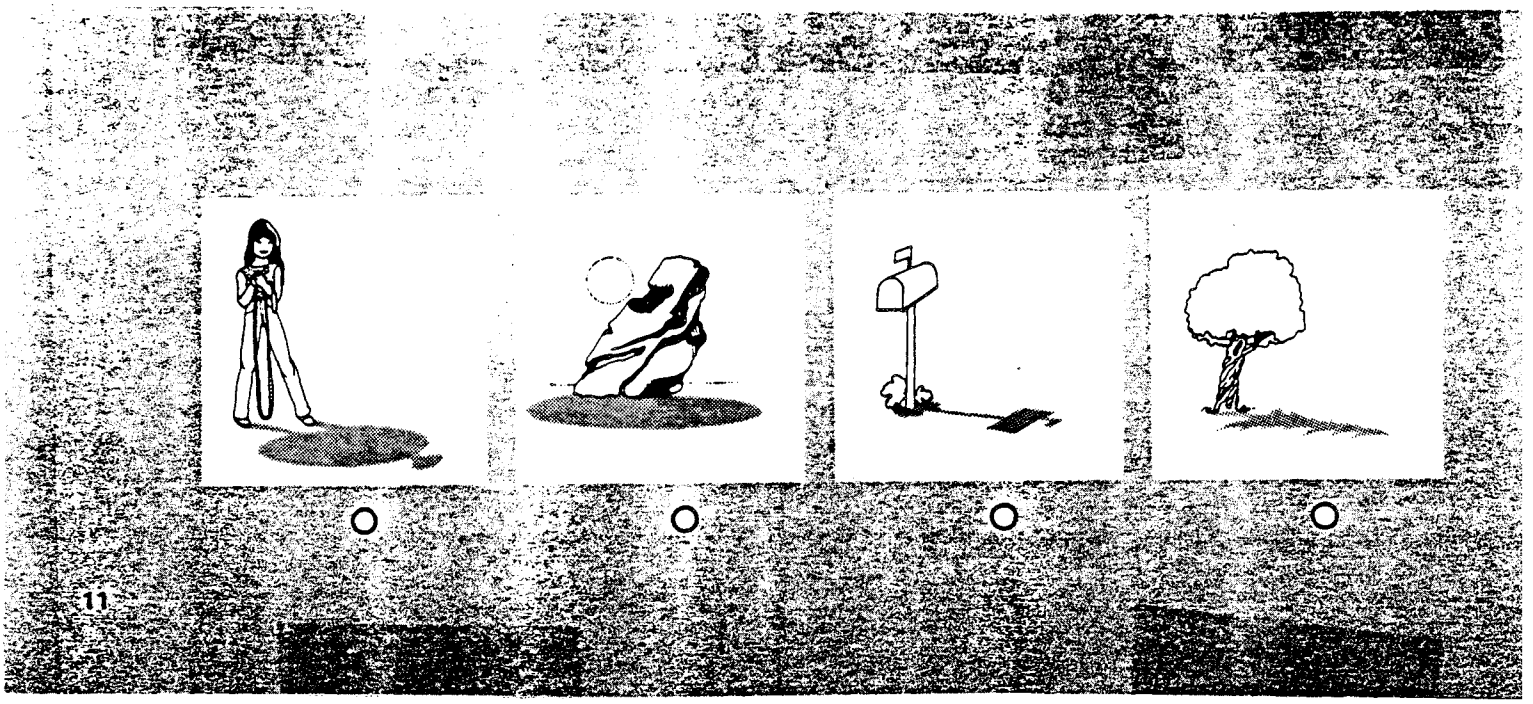
Table 7

Factor Analysis of PTS, 3FM, and MT with Pre- and Poststandardized Tests Using Promax Rotation

Test	Factor 1	Factor 2
WRAT	0.49	0.12
CIRCUS	0.90	-0.09
DRP	0.78	-0.01
TOBE-2	0.11	0.52
STEP	0.81	0.12
PTS	0.21	0.47
3FM	0.46	0.32
MT	-0.04	0.43

Figure 1

TOBE-2 Items



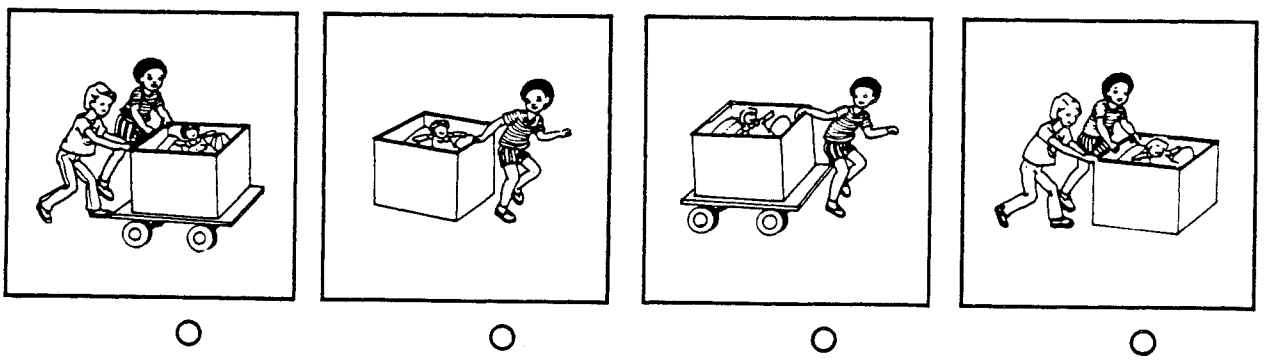
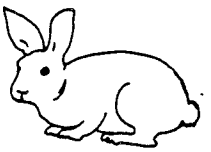


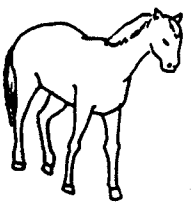
Figure 2

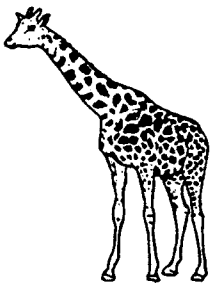
STEP Items

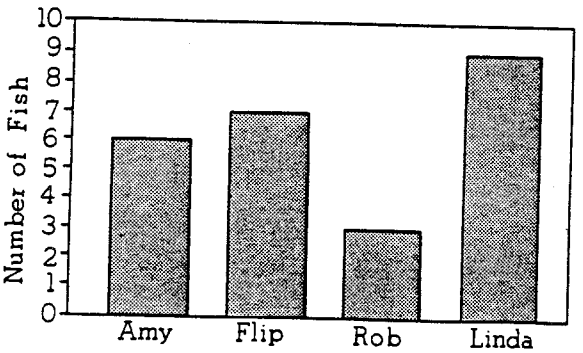
1. When you see these animals at the zoo, which is the tallest?

A) 

B) 

C) 

D) 



7. The graph shows the number of pet fish that four children have. Which child has the most fish?

A) Amy
B) Flip
C) Rob
D) Linda

Figure 3

STEP Items

16. Why does a mother bear growl if a dog comes near her baby?

A) She is hungry.
B) She is afraid the dog will hurt her baby.
C) She wants to play.
D) The dog makes her sleepy.

17. If Tony does not give food and water to his dog every day, the dog will

A) have puppies
B) get sick
C) grow longer hair
D) grow bigger feet

Figure 4

Plant Item

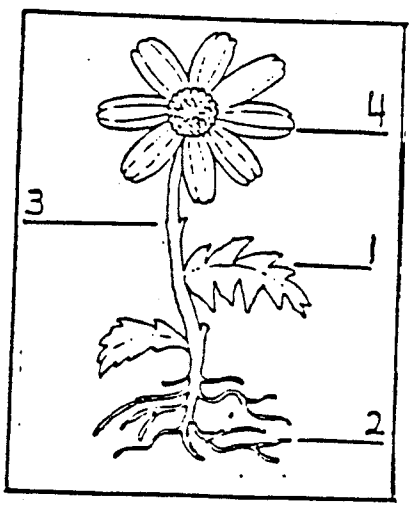


Figure 5

Three Forms of Matter Item

10. PAGE 10. ALL OF THESE CONTAINERS NOW HAVE THE SAME AMOUNT OF WATER. TWO CONTAINERS HAVE LIDS. TWO CONTAINERS DO NOT HAVE LIDS. CIRCLE THE NUMBER OF THE PICTURE OF THE CONTAINER OF WATER THAT WILL EVAPORATE FASTEST.

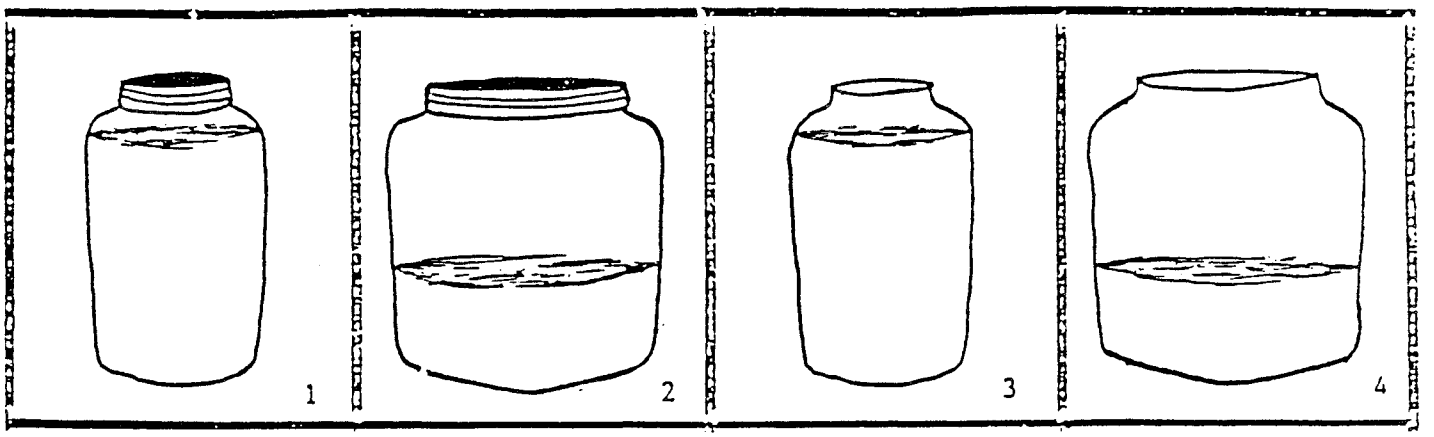


Figure 6

Motion Item

1. LOOK AT PAGE 1. THESE PICTURES SHOW A CHILD ROLLING A BALL TOWARD A WALL. LOOK AT PICTURE 1. TRACE THE ARROW THAT SHOWS WHERE THE BALL WILL GO AFTER IT HITS THE WALL. NOW LOOK AT PICTURE 2. TRACE THE ARROW THAT SHOWS WHERE THAT BALL WILL GO AFTER IT HITS THE WALL. NOW DO THE SAME THING WITH PICTURE 3. TRACE THE ARROW THAT SHOWS WHERE THE BALL WILL GO AFTER IT HITS THE WALL. NOW LOOK AGAIN AT ALL THREE PICTURES. THINK ABOUT WHEN YOU THROW A BALL AGAINST A WALL. THIS CHILD IS GOING TO THROW A BALL AGAINST A WALL. THE BALL WILL BOUNCE BACK EITHER THIS WAY, OR THIS WAY, OR THIS WAY.

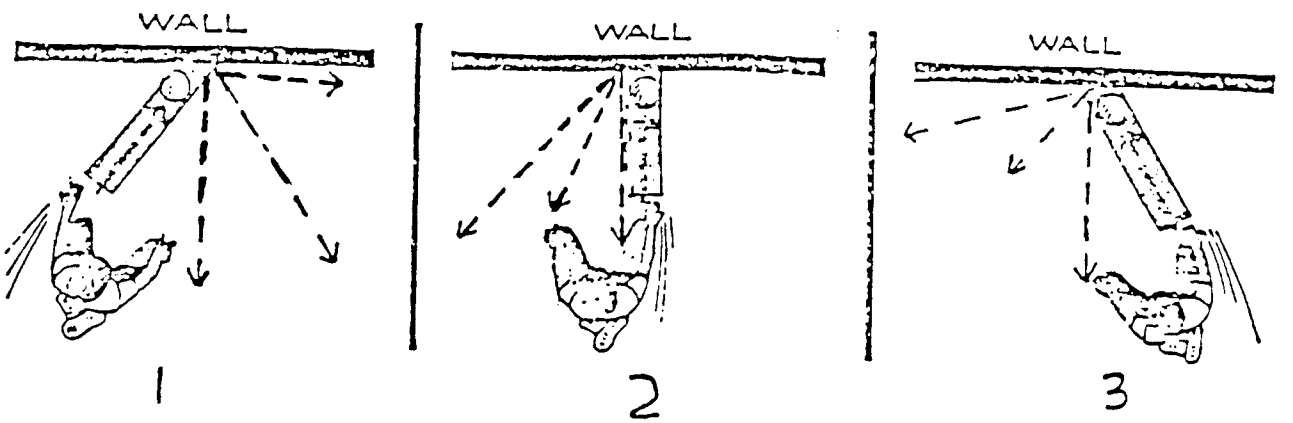


Figure 7

Motion Item

6. A STRAIGHT T LOOKS LIKE THIS. (⊤ Show on board.) IT IS CALLED A STRAIGHT T BECAUSE OF HOW THE LINES GO TOGETHER. SO, THIS T IS STRAIGHT (⅄ draw a straight t on a slant). AND, THIS T IS STRAIGHT (⅄ Draw another straight T on a slant different from the others already drawn on the board.) THIS T IS NOT STRAIGHT. (⅄ Draw a straight T like this on the board.) NOW LOOK AT PAGE ONE IN YOUR BOOKLETS. DRAW A CIRCLE AROUND <u>EVERY</u> T THAT IS A STRAIGHT T.